INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers

Yuan Gong[1], *Sameer Khurana*[1], *Leonid Karlinsky*[2], *James Glass*[1]

[1]MIT CSAIL, USA [2]MIT-IBM Watson AI Lab, USA

{yuangong,glass}@mit.edu

github.com/yuangongnd/whisper-at

Abstract

In this paper, we focus on Whisper [1], a recent automatic speech recognition model trained with a massive 680k hour labeled speech corpus recorded in diverse conditions. We first show an interesting finding that while Whisper is very robust against real-world background sounds (e.g., music), its audio representation is actually not noise-invariant, but is instead highly correlated to non-speech sounds, indicating that Whisper recognizes speech *conditioned* on the noise type. With this finding, we build a unified audio tagging and speech recognition model *Whisper-AT* by freezing the backbone of Whisper, and training a lightweight audio tagging model on top of it. With <1% extra computational cost, Whisper-AT can recognize audio events, in addition to spoken text, in a single forward pass.

1. Introduction

In recent years, significant progress has been made in advancing automatic speech recognition (ASR) performance. Specifically, self-supervised learning schemes such as wav2vec2.0 [2] and Hubert [3] have achieved great success, requiring minimal *labeled* training data. However, since the public model checkpoints are trained with clean speech data (e.g., Librispeech [4] or Libri-light [5]), their robustness in real-world environments is limited. To improve noise robustness, the Whisper [1] model uses 680K hours of *labeled* speech collected from the Internet with *diverse* environments and recording setups as the training data, and reports better robustness over existing ASR models.

In this paper, we first show a counter-intuitive finding that while Whisper is robust against background sounds (noise for ASR), its audio representation is actually not noise-invariant, but instead encodes rich information of non-speech background sounds (shown in Figure 1 and discussed in detail in Section 3), indicating that the Whisper model does not learn a noise-invariant representation, but *encodes* the noise type, and then recognize speech *conditioned* on the noise type.

One exciting application of the above finding is that we can build a *unified* model for ASR and Audio Tagging (i.e., recognize general audio events) based on Whisper since it 1) is robust to noise, and 2) encodes rich general audio event information. Currently, ASR and audio tagging (AT) models are typically performed independently. In many applications such as video transcribing, voice assistants, and hearing aid systems, we desire to get both spoken text and acoustic scene analysis from the audio, but running two systems is computationally expensive. In this work, we show that with <1% extra computational cost, we can make Whisper recognizes audio events together with spoken text in a single forward pass. Our model achieves an mAP of 41.5 on AudioSet, which is slightly worse than standalone AT models, but is nevertheless over 40× faster.

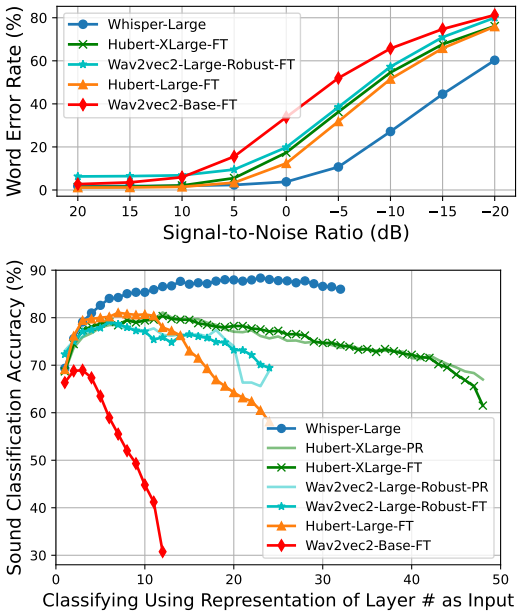

Figure 1: *Surprisingly, the noise robustness of an ASR model correlates* ***positively*** *to the amount of general background sound (noise for ASR) information encoded in their intermediate representations. In the upper figure, we show Whisper is noticeably more robust (smaller word error rate increase) when speech (Librispeech) is contaminated with an increasing amount of background sounds from ESC-50 [6]. In the lower figure, we show the intermediate representations of Whisper lead to the best linear probing sound classification accuracy on the same ESC-50 data, indicating Whisper encodes most background sound information. Unlike other models, Whisper encodes background sound information even in its deepest layer. PR=self-supervised pretrained; FT=PR and fine-tuned model.*

Related Work: To the best of our knowledge, we are the first to report that a robust ASR actually learns a noise-variant representation; most previous work focuses on noise-invariant representations [7, 8, 9, 10, 11]. For ASR and AT model unification, the closest works are [12, 13, 14, 15]. In [12], a unified keyword spotting and audio tagging model is proposed, however, keyword spotting only considers up to 35 words and is a much simpler task than the large-vocabulary continuous speech recognition task we are targeting. In [13, 14], joint ASR and audio tagging/captioning training frameworks are proposed, but in this work, we show that Whisper already encodes rich general audio information even without any explicit audio tagging training. In [15], ASR representations are tested for the audio tagging task, but the overall performance is unsatisfactory.

10.21437/Interspeech.2023-2193

2. Whisper Robust ASR Model

Whisper [1] is a recently proposed robust ASR model that features a standard Transformer [16]-based encoder-decoder architecture. The main novelty of Whisper is not its architecture, but its training data and training scheme. Specifically, the 680K-hour non-public training set contains audio-transcript pairs collected from the Internet with a very broad distribution of audio from many different environments, recording setups, speakers, and languages. Significant effort was made to filter out low-quality data. Compared with the most commonly used Librispeech (960 hours) and Libri-light (60K hours) data that are collected from audiobooks, the Whisper training data is much *larger* and more *diverse*, but also has noisy labels. We identify this as the main factor that differentiates Whisper from existing ASR models. During Whisper training, only text transcripts are used as supervision signals, no audio event labels are given. In this paper, we use the Whisper-Large model unless otherwise stated. Since Whisper is an encoder-decoder model, we only use the audio encoder part of Whisper for audio tagging, which consists of 32 Transformer layers with a dimension of 1280.

3. Noise-Robust ASR Learns Noise-*Variant* Representations

Thanks to the diverse 680K-hour training data, Whisper has been shown to be more robust under white and pub noise than its counterparts [1]. We confirmed this point by evaluating Whisper and other state-of-the-art ASR models on Librispeech clean speech data that were contaminated with ESC-50 [6] environmental sounds with various signal-to-noise ratios (SNRs). As shown in Figure 1 (upper), Whisper has superior performance.

What is the noise-robust mechanism of Whisper? It is commonly believed that the representation of a robust ASR model should be noise-*invariant*, and researchers often set noise-invariance as an explicit inductive bias for robust ASR (e.g., in [7, 8, 9, 10, 11]). However, we, perhaps surprisingly, found that Whisper's representation is actually noise-*variant* and encodes rich non-speech background sound information.

Specifically, we froze the entire Whisper model and input audio samples from the ESC-50 environment sound dataset [6]. We then extracted the intermediate representation from every layer of Whisper and trained a linear layer on top of it to classify the sound class from 50 possible classes. If Whisper did not encode background sound information, or its representations were invariant to background sounds, the sound classification result would be low, and vice versa. As shown in Figure 1 (lower), the Whisper representations had the best ESC-50 sound classification accuracy compared to other SOTA ASR models, indicating that its representation encodes most background sound information. In addition, for all other ASR models, representations from deeper layers led to lower sound classification accuracies, showing that the models are learning to encode speech information, and ignore background sound information. Whisper does not have this behavior, since representations from deeper layers also encode background sound information.

The fact that Whisper is noise-robust while its representation encodes rich background sound information reveals that the robustness mechanism of Whisper is different from other ASR models (including wav2vec2-robust [17]). Instead of learning a noise-invariant representation, it first *encodes* the background sound and then transcribes text *conditioned* on the type of noise. We confirmed this point by further checking the class-wise relationship between Whisper's robustness against a specific background sound class, and its potential ability to recognize the sound class in Figure 2. We found there is indeed a positive correlation between them. Compared to *noise-aware training* [18] that requires manually inputting noise type to the model, Whisper learns it directly from its massive 680K hour training set.

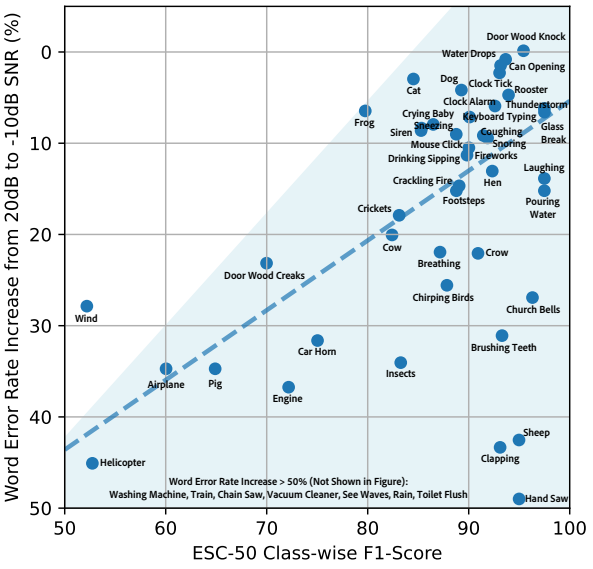

Figure 2: *Class-wise analysis of the relationship between Whisper's robustness against a specific background sound class and its potential ability to recognize the sound. We measure Whisper robustness by its WER increase from clean speech (20dB SNR) to speech contaminated by the specific background sound from ESC-50 (-10dB SNR). The lower the WER increase, the more robust the model (Y-axis). We estimate the potential ability of Whisper to recognize the sound by training a linear layer on top of the Whisper encoder's last-layer representation for the sound classification task on the same ESC-50 dataset (without speech mixed-in, the Whisper model is frozen) and show the class-wise F1-score. The higher the F1-score, the better Whisper can potentially recognize the sound class (X-axis). Blue dashed line: we observe a positive correlation between Whisper's robustness against a background sound type and its potential ability to recognize it. Blue shading: we observe most sound classes lie in the right-bottom triangle area, indicating that Whisper is not robust to the type of sound if it cannot recognize the sound type. Right-bottom outliers: there are some background sounds that Whisper can potentially recognize but is not robust to, which is expected as some noises heavily overlap with the speech and are impossible to be robust to. In short, we find the potential ability to recognize a sound type is a necessary but not sufficient condition for Whisper to be robust to it.*

Note that the discussion in this section is mostly based on Whisper, and our experiments do not indicate that noise-invariance does not help noise-robust ASR, nor that a noise-robust ASR's representation should be noise-variant. In fact, we believe encouraging noise-invariant representations [7, 8, 9, 10, 11] is a practical solution in self-supervised learning or small data cases. Whisper training requires industry-level computational resources and is expensive. What we hope to convey is that a noise-robust ASR model does not have to learn a noise-invariant representation, and that there exist other ways to be noise-robust - a noise-conditioned model like Whisper can, and indeed does, work very well.

4. Unifying ASR and Audio Tagging Model

One exciting application of the finding in Section 3 is that we are able to build a *unified* model for ASR and Audio Tagging based on Whisper to recognize spoken text and background sounds (e.g., music, horn, etc) simultaneously, which is highly desirable in applications such as video transcribing, voice assistants, and hearing aid systems. Whisper is ideal as a backbone for such a unified model because 1) it is robust to background sounds, and 2) its intermediate representations encode rich general audio event information, which serves as a solid base for audio tagging. Nonetheless, the original Whisper does not output sound labels, so we need to train a model on top of Whisper intermediate representations to enable it to predict a sound class. Note that we intentionally do not modify the original weights of the Whisper model, but instead add new audio tagging layers on top of it so that the Whisper ASR ability is not changed and text and audio labels can be generated in a *single* forward pass. We call this unified ASR and Audio Tagging model *Whisper-AT*.

In previous sections, we applied a basic linear layer on the representation of a single layer for probing purposes. In this section, we discuss more advanced methods that lead to better audio tagging performance.

1. **`Last-MLP`**: The most basic method, we first apply a temporal mean pooling over the last layer representation of Whisper and then apply a linear layer to map it to the prediction.
2. **`WA-MLP`**: As shown in Figure 3, we find the last layer is not optimal for all sound classes. Thus we weighted average (WA) the representations from all layers and set the weight to be learnable before temporal mean pooling and linear layer, so this approach leverages representations from all layers.
3. **`WA-Tr`**: Temporal mean pooling removes all temporal details, and a single linear layer may be too simple for audio tagging. Therefore, we replace the linear layer of `WA-MLP` with a single-head temporal Transformer layer for this model.
4. **`TL-Tr`**: Time and layer-wise Transformer (our main method, shown in Figure 4). Though weighted averaging leverage representation of all layers, all sound classes use a *fixed* set of weights. In Figure 3, we show that different sound classes achieve their best performance using different representation layers. Therefore, ideally, each class should have its own set of weights. This motivates us to build an attention mechanism over the *layers*. Specifically, we apply another layer-wise Transformer to the output of the temporal Transformer.

Efficient Design: As the original goal of Whisper-AT is being more computationally efficient than two independent ASR and AT models, we aim to minimize the extra cost for audio tagging. Introducing a new Transformer layer in `WA-Tr` and `TL-Tr` is relatively expensive. Consider the complexity of Transformer is $O(d^2n + dn^2)$, where d is the dimension and n is the input length of the Transformer, for each 10-second input audio, the representations of each Whisper layer is in the shape of (n=500, d=1280). If the temporal and layer Transformer have the same n and d as Whisper, their computational cost is not negligible. Therefore, as illustrated in Figure 4, we propose the following efficient design: 1) We add a mean pooling layer to each Whisper representation to lower the time sequence length n from 500 to 25; 2) We add an optional linear projection layer to lower d from 1280 to 512 before audio tagging Transformers (denoted by $\texttt{TL-Tr}_{512}$); and 3) For `WA-Tr`, we first conduct weighted averaging and then apply a temporal Transformer, for `TL-Tr`, we use a single temporal Transformer for all layers. Thus both `WA-Tr` and `TL-Tr` only need one temporal Transformer.

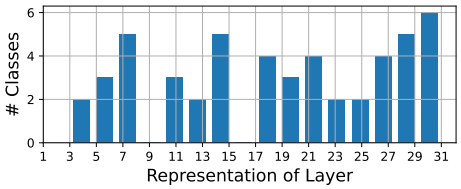

Figure 3: *Histogram of the best Whisper representation layer (1-32) for the 50 ESC-50 sound classes. We train a linear layer on top of the representation of each of the 32 Whisper layers for ESC-50 sound classification, compute the class-wise F1-Score, and find the best representation layer for each sound class. Different sound classes get the best F1-score on representations of different layers.*

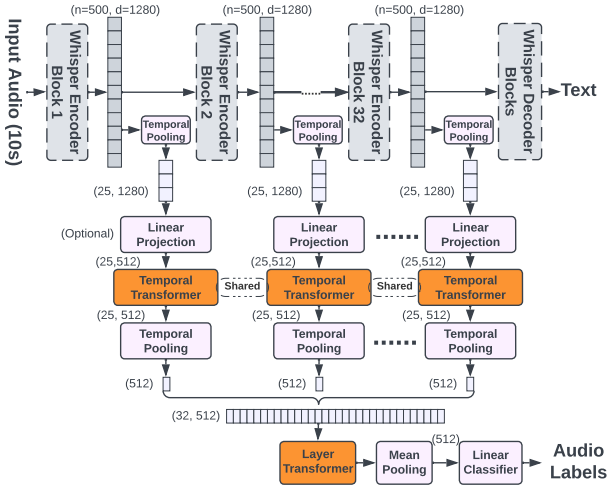

Figure 4: *The proposed time and layer-wise Transformer model.*

5. Experiments

As mentioned in Section 4, we intentionally freeze the weights of the original Whisper model, so the ASR performance of Whisper-AT is exactly the same as the original Whisper [1]. Thus we only conduct experiments on the audio tagging task.

5.1. Experiment Settings

Dataset: We use AudioSet and ESC-50 datasets following standard evaluation protocols. AudioSet [20] is a collection of over 2 million 10-second audio clips excised from YouTube videos and labeled with the sounds that the clip contains from a set of 527 labels. We train our model with both the balanced training set (AS-20K) and full training set (AS-2M) and report mAP on the evaluation set. ESC-50 [6] consists of 2,000 5-second environmental audio recordings organized into 50 classes; we evaluate our model using the official 5-fold cross-validation protocol.
Hyper-Parameters: We use the standard training pipeline in prior AT work [21, 22, 26, 27]. For all experiments, we use a batch size of 48 and an Adam optimizer [28]. For the proposed $\texttt{TL-Tr}_{512}$ model, we use an initial learning rate of 2e-4, 1e-4, and 5e-4, and train the model for 30, 5, and 30 epochs for AS-20K, AS-2M, and ESC-50, respectively. For baseline methods, we search the learning rate to ensure a fair comparison.

5.2. Experiment Results

We show the main results in Table 1. The key conclusions are:

Table 1: *Audio tagging performance comparison on AS-20K, AS-2M (mAP), and ESC-50 (accuracy).* †*ASR backbone parameters and FLOPs are not included.* ⋆*Speed-up = 1/FLOPs, compared with AST; FLOPs computed by* `fvcore` *[19].* ‡*: labeled AS-2M data is also used.* ⋆⋆ *AS-2M experiment is expensive, we skip it when AS-20K and ESC50 experiments already shown clear differences. End-to-End fine-tuning results are shown in grey text as the comparison is not exactly fair.*

Model	Training Setting	Method	AS-20K	AS-2M	ESC-50	AT #Params†	AT Speed-Up†⋆
Existing Standalone Audio Tagging Models							
AudioSet Baseline [20]	Fine-Tuning	End-to-End	-	31.4	-	-	-
AST [21]	Fine-Tuning	End-to-End	34.7	45.9	88.8	87M	1 × (133G FLOPs)
SSAST [22]	Fine-Tuning	End-to-End	31.0	-	88.7	87M	1 ×
PANNs [23]	Fine-Tuning	End-to-End	27.8	43.9	94.7‡	81M	2.5 ×
MAE-AST [24]	Fine-Tuning	End-to-End	30.6	-	90.0	87M	2.7 ×
Audio-MAE [25]	Fine-Tuning	End-to-End	37.0	47.3	94.1	87M	2.7 ×
Existing Automatic Speech Recognition Models							
Hubert X-Large [3]	Frozen	WA-MLP	18.5	- ⋆⋆	82.2	0.7M	195K ×
Hubert X-Large [3]	Frozen	$\texttt{TL-Tr}_{1280}$	20.2	-	83.6	40M	5 ×
wav2vec2-Large-Robust [17]	Frozen	WA-MLP	18.1	-	78.5	0.5M	244K ×
wav2vec2-Large-Robust [17]	Frozen	$\texttt{TL-Tr}_{1024}$	20.2	-	82.8	26M	17 ×
Whisper-AT							
Whisper-Large	Frozen	Last-MLP	20.6	20.3	87.0	0.7M	195K ×
Whisper-Large	Frozen	WA-MLP	25.7	32.4	90.2	0.7M	195K ×
Whisper-Large	Frozen	WA-Tr	32.1	41.0	91.0	20M	270 ×
Whisper-Large	Frozen	$\texttt{TL-Tr}_{1280}$	33.0	42.1	91.1	40M	8 ×
Whisper-Large	Frozen	$\texttt{TL-Tr}_{512}$	32.8	41.5	91.7	7M	42 ×
Whisper-Large	Fine-Tuning	End-to-End	34.7	45.7	90.0	655M	0.4 ×
Whisper-Small	Fine-Tuning	End-to-End	31.9	44.1	88.9	94M	2.5 ×

First, Whisper-AT is significantly stronger than Hubert X-Large [3] and wav2vec2-Large-Robust [17] on audio tagging, demonstrating that Whisper is not only the most robust ASR model but also the strongest audio tagging backbone.

Second, comparing the four Whisper-AT models, the proposed `TL-Tr` model leads to the best performance with higher computational overhead. However, by projecting the Transformer dimension from 1280 to 512, $\texttt{TL-Tr}_{512}$ strikes a balance between performance and efficiency, as its FLOPs are less than 1% of the Whisper ASR FLOPs yet it performs almost the same as $\texttt{TL-Tr}_{1280}$. In Table 2, we further study the relationship between the audio tagging performance and Transformer dimension d for `TL-Tr`. Even $\texttt{TL-Tr}_{128}$ provides reasonably good audio tagging performance, while its computational cost is almost free (<0.1% FLOPs of the Whisper ASR FLOPs).

Third, Whisper-AT is slightly worse than SOTA standalone audio tagging models but is much more efficient. The proposed $\texttt{TL-Tr}_{512}$ achieves 32.8 mAP, 41.5 mAP, and 91.7 accuracy on AS-20K, AS-2M, and ESC-50, respectively, and is 42 times faster and 11 times smaller than AST [21]. Note that we target the cases that the user is already running an ASR and want to get additional audio labels, so we only compare the *additional* cost for AT and do not include the cost of ASR in this comparison.

Fourth, how does Whisper perform in the end-to-end fine-tuning setting, and how does it compare to SOTA audio tagging models? We add a new Transformer layer on top of the Whisper encoder and train the entire model end-to-end (new layer uses a 10-100× larger learning rate). For a fair comparison, we also test Whisper-Small which is of similar size to SOTA audio tagging models. We find Whisper-Small performs similarly with previous self-supervised pretrained models such as SSAST [22] and MAE-AST [24] after fine-tuning.

Finally, we test the audio tagging performance of smaller Whisper models. As shown in Figure 5, smaller models have weaker audio tagging performance but the difference between Whisper-Small, Medium, and Large is minor. We also test the ASR noise-robustness of these models on speech contaminated by ESC50 background sounds; larger models are more robust. We again observe a positive correlation between ASR noise robustness and AT performance. In addition, Whisper-Base (74M parameters) is already more robust in ASR and stronger in audio tagging than Hubert-X-Large (964M parameters).

Table 2: *Performance and efficiency impact of* `TL-Tr` *Transformer dimension d.*

Tr Dim d	FLOPs (G)	#Params (M)	AS-20K	ESC-50
128	0.31	0.6	30.0	91.4
256	0.94	2.1	32.0	92.0
512	3.17	7.2	32.8	91.7
768	6.72	15.6	33.0	91.4
1280	16.42	40.0	33.0	91.1

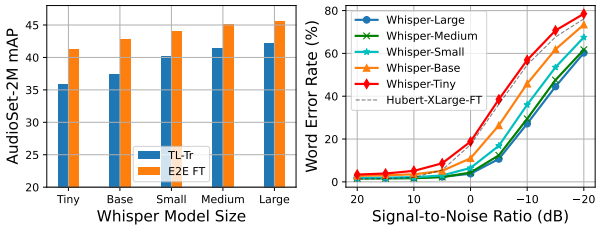

Figure 5: *AS-2M audio tagging performance (left) and ASR robustness (right) of the Whisper model family.*

6. Conclusion

The Whisper ASR model revives the supervised learning scheme by using a massive and diverse training corpus. In this paper, we report an intriguing property of Whisper that while being very robust, the audio representation of Whisper is actually noise-variant and encodes rich background sound information. Based on this finding, we propose a unified audio tagging and ASR model called *Whisper-AT*. With only less than 1% additional cost, Whisper-AT can recognize the background sound in addition to spoken text in a single forward pass.

Acknowledgments: This research is supported by the MIT-IBM Watson AI Lab.

7. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[5] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[6] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[7] M. Van Segbroeck *et al.*, "Unsupervised learning of time–frequency patches as a noise-robust representation of speech," *Speech Communication*, vol. 51, no. 11, pp. 1124–1138, 2009.

[8] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, "Invariant representations for noisy speech recognition," *NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop*, 2016.

[9] A. Sriram, H. Jun, Y. Gaur, and S. Satheesh, "Robust speech recognition using generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5639–5643.

[10] D. Liang, Z. Huang, and Z. C. Lipton, "Learning noise-invariant representations for robust speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 56–63.

[11] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, M.-H. Wu, X. Fang, and L.-R. Dai, "A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3174–3178.

[12] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "Unikwat: Unified keyword spotting and audio tagging," *arXiv preprint arXiv:2209.11377*, 2022.

[13] N. Moritz, G. Wichern, T. Hori, and J. Le Roux, "All-in-one transformer: Unifying speech recognition, audio tagging, and event detection." in *INTERSPEECH*, 2020, pp. 3112–3116.

[14] C. Narisetty, E. Tsunoo, X. Chang, Y. Kashiwagi, M. Hentschel, and S. Watanabe, "Joint speech recognition and audio captioning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7892–7896.

[15] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[17] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proc. Interspeech 2021*, 2021, pp. 721–725.

[18] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7398–7402.

[19] "vcore," https://github.com/facebookresearch/fvcore.

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[21] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[22] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.

[23] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[24] A. Baade, P. Peng, and D. Harwath, "MAE-AST: Masked Autoencoding Audio Spectrogram Transformer," in *Proc. Interspeech 2022*, 2022, pp. 2438–2442.

[25] P.-Y. Huang, H. Xu, J. B. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," in *Advances in Neural Information Processing Systems*, 2022.

[26] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[27] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 200–14 213, 2021.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.